DOCUMENT RESUME

ED 348 378 TM 018 743

AUTHOR Kromrey, Jeffrey D.; Bacon, Tina P.

TITLE Item Analysis of Achievement Tests Based on Small Numbers of Examinees.

SPONS AGENCY Florida State Dept. of Education, Tallahassee.; University of South Florida, Tampa. Inst. for Instructional Research and Practice.

PUB DATE Apr 92

NOTE 45p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Achievement Tests; Comparative Analysis; Difficulty Level; *Estimation (Mathematics); *Item Analysis; Mathematical Models; Monte Carlo Methods; Psychometrics; *Sample Size; *Statistical Bias; Test Construction; *Test Items

IDENTIFIERS Alpha Coefficient; Item Discrimination (Tests); Point Biserial Correlation

ABSTRACT

A Monte Carlo study was conducted to estimate the small sample standard errors and statistical bias of psychometric statistics commonly used in the analysis of achievement tests. The statistics examined in this research were: (1) the index of item difficulty; (2) the index of item discrimination; (3) the corrected item-total point-biserial correlation coefficient; and (4) coefficient alpha. Sample sizes of 5, 10, 20, 40, 80, and 160 were evaluated. One thousand samples of each size were drawn with replacement from each of 10 archival data files from teacher subject area tests. These files represent pseudo-populations whose parameters are directly calculable and from which the sampling bias and errors of statistics are empirically estimable. The behavior of each statistic was evaluated by computing the standard error of the statistic for each sample size and each pseudo-population, and by computing the statistical bias of the statistic for each sample size and each pseudo-population. Results are interpreted in terms of their applications to test development. Nine tables present study data, and nine figures illustrate the discussion. There is a 13-item list of references. (Author/SLD)

Item Analysis of Achievement Tests Based on Small Numbers of Examinees

Jeffrey D. Kromrey

University of South Florida

Tina P. Bacon

University of South Florida

This research was supported in part by a grant from the Florida Department of Education and the Institute for Instructional Research and Practice at the University of South Florida.

RUNNING HEAD: Item Analysis

Paper presented at the annual conference of American Educational Research Association, April 19-24, 1992, San Francisco, CA.

Abstract

A Monte Carlo study was conducted to estimate the small sample standard errors and statistical bias of psychometric statistics commonly used in the analysis of achievement tests. The statistics examined in this research were (a) the index of item difficulty, (b) the index of item discrimination, (c) the corrected item-total point-biserial correlation coefficient, and (d) coefficient alpha. Sample sizes of 5, 10, 20, 40, 80, and 160 were evaluated. One thousand samples of each size were drawn with replacement from each of ten archival data files from teacher subject area tests. Results were interpreted in terms of applications to test development.

Item Analysis of Achievement Tests Based on Small Numbers of Examinees

The traditional techniques of item analysis (i.e., the calculation of item difficulty indices, item discrimination indices, and distractor analyses) may have limited utility when the number of examinees on which the calculations are based is small. Two statistical issues to consider in the application of these techniques to small samples are (a) the magnitudes of the standard errors of the statistics, and (b) the potential for statistical bias in the estimation of population parameters.

The purpose of this research was to develop estimates of the standard errors and biases of item difficulty, discrimination indices, and coefficient alpha when the calculations are based on small samples of examinees.

Knowledge of the standard errors of statistics used in the analysis of achievement test items is valuable for the interpretation of the results of an item analysis. For example, the width of a confidence interval around a calculated index of discrimination provides information about the expected amount of variation in the obtained magnitude of that statistic under repeated sampling.

With the exception of the item difficulty index, the statistics used in item analysis do not provide easily calculated standard errors (Perry & Michael, 1954). Further, asymptotic

formulas for standard errors that provide useful approximations for large sample sizes are frequently inaccurate when applied to small samples.

The standard errors of item statistics are typically ignored in traditional item analyses. With large samples of examinees, the practice of ignoring statistical errors is probably acceptable because the magnitudes of the standard errors are reasonably small in such circumstances (being inversely related to the number of observations on which the statistics are calculated). With small numbers of examinees, however, the practice of ignoring standard errors should be seriously questioned.

Rationale

The classical true score model for computing item and test score indices has served test constructors well for many years. The simplicity of the model and the reasonable ease of computing indicators of test or item functioning are some of the model's advantages. There are, however, some concerns that arise in the use of traditional methods for test construction and revision.

The value of item indices, such as difficulty and discrimination, are group-dependent (Hambleton, 1989). The computed value of these statistics vary according to the attributions, skills, or ability level of the derivation sample. A group of examinees possessing greater ability will result in higher item difficulty indices than when the statistic is

computed from performance of lower ability examinees.

Likewise, discrimination and reliability indices are impacted by the group variability. The magnitude of the correlation coefficient is dependent on the homogeneity or heterogeneity of the group, with diverse samples having higher values than groups having similar ability levels (Lord & Novick, 1968).

It is, in part, the nature of the indices to be group-dependent that raises the question of what sample size is necessary to provide the test constructor confidence in using the indices for test refinement and describing test functioning. Many researchers have proposed rules-of-thumb for determining sample size for conducting item analysis. Nunnally (1967) suggested that the test constructor have 5 to 10 times as many subjects as items. Crocker and Algina (1986) proposed that sample sizes of 200 subjects would offer reasonable statistical stability. Other researchers and commercial test developers have recommended sample sizes ranging from 300 to 3000 depending upon the target population to be served by the instrument (Conrad, 1948; Henryson, 1971; and Swineford, 1974).

Little empirical work has been done that directly investigated the impact of small sample sizes on standard error and statistical bias of traditional test and item indices. In fact, the majority of the research has focused on sample size and sample variability in the application of item-response methods to

test construction and equating.

One study conducted by Nevo (1980) investigated the effects of various sample sizes on the accuracy of rank-ordering or categorizing of traditional item indices. Nevo's findings suggested that sample sizes as low as 100 may be sufficient if the researcher's goal is to position items in relation to each other. But the question of the absolute value or loss of accuracy for individual item indices estimated from small samples was not addressed.

The lack of empirical efforts examining the relationship of sample size to standard error and bias in estimating traditional item indices suggests the need for such study. Empirical findings may even provide the opportunity to suggest sample size guidelines for different indices and the estimated cost to accuracy and utility in making decisions based on values obtained using various sample sizes from a defined population.

Method

Four item analysis statistics were examined in this study: (a) coefficient alpha (equivalently, the Kuder-Richardson Formula 20), (b) the item difficulty index (the proportion of examinees responding correctly to the test item), (c) the item discrimination index (the difference between the item difficulty index for the top 27% of the examinees and the item difficulty index for the bottom 27% of the examinees), and (d) the item-total point-biserial correlation coefficient (the correlation

between performance on the test item and total test score, corrected for overlap).

This research was conducted by drawing random samples from existing archival examination data files. These files represent pseudo-populations whose parameters are directly calculable, and from which the sampling bias and errors of statistics are empirically estimable. Data files from ten teacher subject area examinations were used. The tests used in this research are listed in Table 1.

Insert Table 1 about here

From each pseudo-population, small random samples were drawn with replacement. One thousand samples each of sizes 5, 10, 20, 40, 80, and 160 records were drawn from each of the ten pseudo-populations. In each sample, the indices of item difficulty and discrimination and the item-total point-biserial correlation were computed for each item. In addition, the value of coefficient alpha for the test was computed for each sample of examinees.

The behavior of each statistic was evaluated by (a) computing the standard error of the statistic for each sample size and each pseudo-population, and (b) computing the statistical bias of the statistic for each sample size and each pseudo-population.

The standard error of each statistic was computed as the standard deviation of the sample estimates of the statistic about

the mean value of the statistic:

$$SE_{\theta j} = ([\Sigma(\hat{\theta}_{ij} - \hat{\theta}_{.j})^2]/(N-1))^{\frac{1}{2}}$$

where

$SE_{\theta j}$ = standard error of the item statistic in samples of size j,

$\hat{\theta}_{ij}$ = value of the statistic computed from sample i of size j,

$\hat{\theta}_{.j}$ = mean value of the statistic in the 1000 samples of size j.

The bias of each statistic was computed as the difference between the mean value of the statistic in the 1000 samples and the value of the statistic in the pseudo-population:

$$Bias_{\theta j} = \hat{\theta}_{.j} - \theta$$

where

$Bias_{\theta j}$ = statistical bias in the estimation of the statistic in samples of size j,

$\hat{\theta}_{.j}$ = mean value of the statistic in the 1000 samples of size j,

θ = value of the statistic in the pseudo-population.

All program code for the random sampling and statistical computations was written in SAS, Version 6.06.

Results

Because the results of the statistical bias analyses and standard error analyses were quite consistent across the ten subject area tests examined in this research, and to conserve space, detailed results are presented for only one subject area. Additional detailed results are available from the authors.

Index of Item Difficulty

Statistical Bias. Little statistical bias is evident in the estimation of the item difficulty index, even with samples as small as size 5. Box-and-whisker plots of item-level statistical bias for each sample size included in the study are presented in Figure 1. To construct this plot, the difference between the average p-value of each item (computed across the 1000 samples) and the p-value calculated from all examinees in the pseudo-population was computed. The plot provides the distribution of these differences for each test item on the test form. As is evident in Figure 1, the expected value of this difference is nearly zero for each sample size examined. Although the variability of the individual item biases decreases with increasing sample size, even with samples of size 5, the item biases range from only -0.02 to 0.015.

Insert Figure 1 about here

To further explore statistical bias in the estimation of item p-values, items were grouped according to the magnitude of the p-value obtained from the pseudo-population. The average bias within each group for each sample size was computed. The results of this analysis are presented as Table 2. No systematic relationship between the p-value of the item in the pseudo-population and the degree of statistical bias is evident in this table. Most importantly, in all categories, the magnitude of bias is negligible.

Insert Table 2 about here

Standard Error. As expected, the standard error of the item difficulty index is related to the value of the item difficulty (being largest at p=0.5). Table 3 presents, for each sample size examined, the average standard error for items grouped according to item difficulty. The relationship between the magnitude of the item difficulty index and its standard error is evident in this table. In samples of size 5, the average standard error of items ranging from p=.40 to p=.59 is 0.22, while items with p>.90 present an average standard error of less than 0.10 and items with p<.10 present an average standard error of 0.13. Note that

the standard error curve flattens with larger sample sizes (this effect is best seen in the graphic presentation of these data in Figure 2). With samples of size 20, the standard errors ranged from 0.04 to 0.11. At N=40, the average standard error for items in the middle of the range is less than 0.08, while the average standard errors for the extreme values of p are between 0.03 and 0.04. Finally, at N=160, the average standard error ranges only from 0.01 to 0.04.

Insert Table 3 & Figure 2 about here

Index of Item Discrimination

Statistical Bias. In contrast to the results obtained with the item difficulty index, substantial statistical bias is evident in the estimation of the item discrimination index when small samples of examinees are used. Box-and-whisker plots of item-level statistical bias for each sample size included in the study are presented in Figure 3. As with the plots of the item difficulty index, this plot presents the distribution of differences between the average D-value of each item (computed across the 1000 samples) and the D-value for the item computed from the pseudo-population. As is evident in Figure 3 the expected value of this difference is substantially less than zero for samples of size 5 and 10. The average bias in estimating the item discrimination index is approximately -0.1 for samples of

size 5. The middle fifty percent of the items on the examination form present biases ranging from -0.04 to -0.15. For samples of size 10, the average bias in the estimation of item discrimination is reduced to -0.04, and the middle fifty percent of the items present biases ranging from -0.02 to -0.06. With samples of size 20 or larger, the average bias is reduced to a negligible level, a result which was consistent across the ten examination forms included in this study.

Insert Figure 3 about here

To further explore the statistical bias in the estimation of item D-values, the test items were grouped according to the magnitude of the D-value obtained from the pseudo-populations. The average bias within each group for each sample size was computed. The results of this analysis are presented in Table 4 and Figure 4.

Insert Table 4 & Figure 4 about here

The degree of statistical bias in the estimation of the item discrimination index is proportional to the population value of the discrimination index. With samples of size 5, the average bias for items with discrimination indices less than 0.10 is -0.016. In contrast, the average statistical bias for items with

discrimination indices between 0.30 and 0.39 is -0.117, while highly discriminating items (0.70 to 0.79) present an average bias of -0.242. The graph of statistical bias by population value of the index shows a nearly linear relationship between the extent of the bias and the value of the statistic in the pseudo-population. The negative bias in the estimation is substantially reduced in samples of size 20 or larger.

Standard Error. The average standard error of the item discrimination index for each sample size examined is presented in Table 5. The standard errors for the item discrimination index are notably larger than the errors evident for the indices of item difficulty. In samples of size 5, the average standard error of items ranging from D values of 0.30 to 0.39 is 0.45, while items with D<.10 present an average standard error of 0.27. At N=20, the standard errors of the item discrimination index range from 0.15 to 0.28, and only at sample sizes of 160 do the standard errors across the range fall below 0.10. Standard error curves for each sample size are presented in Figure 5.

Insert Table 5 & Figure 5 about here

Item-Total Point Biserial Correlation

Statistical Bias. The analysis of statistical bias in the estimation of the item point biserial correlation yielded similar results to the analysis of the bias in the estimation of the item

discrimination index, although the magnitude of the small sample statistical bias is substantially reduced. Box-and-whisker plots of item-level statistical bias for each sample size included in the study are presented in Figure 6. As with the previous presentations, this plot presents the distribution of differences between the average value of the point biserial correlation for each item (computed across the 1000 samples) and the value of the point biserial correlation computed from the pseudo-population. The small sample bias is evident in Figure 6, as the expected value of this difference is substantially less than zero for samples of size 5 and 10. The average bias in estimating the item-total point biserial correlation is approximately -0.06 for samples of size 5. The middle fifty percent of the items on the examination form present biases ranging from -0.02 to -0.08. For samples of size 10, the average bias in the estimation of the item-total point biserial correlation is reduced to -0.02, and the middle fifty percent of the items present biases ranging from zero to -0.03. With samples of size 20 or larger, the average bias is reduced to a negligible level.

Insert Figure 6 here about here

To further explore statistical bias in the estimation of the item-total point biserial correlation, items were grouped according to the magnitude of the point biserial correlation

obtained from the pseudo-population. The average statistical bias within each group for each sample size was computed. The results of this analysis are presented in Table 6 and Figure 7. As with the estimation of the item discrimination index, the degree of statistical bias is related to the population point biserial correlation. With samples of size 5, the average bias for items with discrimination indices less than 0.10 is -0.02. In contrast, the average bias for items with discrimination indices between 0.20 and 0.29 is -0.058, while highly discriminating items (0.40 to 0.49) present an average bias of -0.064.

Insert Table 6 & Figure 7 about here

Interestingly, the bias for items with values of the point biserial correlation between 0.50 and 0.59 showed a slightly reduced level of bias (-0.056). The highest degree of statistical bias was obtained for items with point-biserial correlations between 0.30 and 0.39. The graph of statistical bias by population value of the index shows the u-shaped relationship between the extent of the bias and the value of the statistic in the pseudo-population. This u-shaped relationship was found in six of the subject area tests examined in this study, while in the remaining four subject area tests, the relationship between degree of bias and the population value of the statistic was nearly linear. As with the estimation of the

item discrimination index, the negative bias in the estimation of the point biserial correlation was substantially reduced in samples of size 20 or larger.

Standard Error. The standard errors of the point biserial correlation for each sample size examined is presented in Table 7. The standard errors for the point biserial correlation are similar in magnitude to those obtained for the item discrimination index. In samples of size 5, the average standard error of items ranging from r_{pbis}=0.20 to r_{pbis}=0.39 is 0.43, while items with r_{pbis}<0.10 present an average standard error of 0.31. At N=20, the standard errors of the point biserial correlation range from 0.23 (for r_{pbis} between 0.10 and 0.19) to 0.17 (for r_{pbis} between 0.50 and 0.59). In general, across the ten examination included in this study, the standard errors for the point biserial correlation were smaller than those obtained for the discrimination index, although the difference was negligible. As with the index of item discrimination, all of the average standard errors for the point biserial correlation fall below 0.10 with samples of size 160, although at N=80, the standard errors are very close to this value. The standard error curves for each sample size are presented in Figure 8.

Insert Table 7 & Figure 8 about here

Coefficient Alpha

Statistical Bias. The analysis of statistical bias in the estimation of coefficient alpha is presented in Table 8. This table presents the bias in estimation of alpha for each test form and each sample size examined in this research. All of the sample estimates of alpha present negative bias, although with the large sample sizes, the extent of the statistical bias is trivial.

Insert Table 8 about here

With samples of size five, the magnitudes of bias ranged from -0.059 (test form 9) to -0.126 (test form 3). Doubling the sample size to samples of size 10 reduced bias to the range of -0.019 (test forms 6 and 9) to -0.091 (test form 3). With samples of size 20, the statistical bias in the estimation of coefficient alpha was less than -0.05 for all ten test forms examined, and for samples of size 40 the bias was less than -0.025.

Standard Error. The standard errors of alpha, estimated for each of the ten examinations, are presented in Table 9. The average standard error ranged from 0.02, for samples of size 160, to 0.18 for samples of size 5. For samples of size 20, the average standard error of coefficient alpha was 0.07 and only one of the ten examinations showed a standard error greater than

0.10. Dropping to samples of size 10 increased the average standard error to 0.12, and seven of the ten examinations showed standard errors greater than 0.10. Box-and-whisker plots of the distributions of the samples of coefficient alpha are presented in Figure 9.

Insert Table 9 & Figure 9 about here

Discussion

Of the statistics examined in this research, only the index of item difficulty provided unbiased estimates of the population value across the breadth of sample sizes examined. However, the biases evidenced in the item discrimination index, the item-total point-biserial correlation coefficient, and coefficient alpha were substantially reduced with samples of size 20 or larger. The negative biases obtained for the item discrimination index and the item-total point-biserial correlation coefficient were related to the population magnitudes of the statistics, with greater degrees of statistical bias being associated with more discriminating items. Of the two statistics, the item-total point-biserial correlation was the superior performer with small samples, showing about half the degree of the bias as the item discrimination index. The standard errors for these two statistics were comparable, but both showed substantially larger standard errors than those obtained for the item difficulty

index.

The practical implications of these results are twofold. First, the results support the use of the sample estimates of the item difficulty index even with small samples, provided that the standard errors are considered in their interpretations. Fortunately, the standard errors are considerably reduced at the extreme values of item difficulty. In pilot testing operations, items with extremely high or low values of difficulty are likely to be flagged for further examination, possible deletion or modification. The availability of greater precision of estimation at the extreme values increases confidence in data support for such decisions.

Secondly, the statistical bias that is evident in the sample estimates of the item discrimination index and the point-biserial correlation coefficient suggest greater caution in their interpretation when the number of examinees is small. In addition, the standard errors of these estimates are so large that a conservative interpretation (i.e., using a two standard error confidence band) renders the estimates virtually useless because of their lack of precision.

The performance problems evidenced in small samples by both the point biserial correlation and the discrimination index suggest the need for an alternative index of discrimination. Because the point biserial correlation is statistically related to the independent-means t-test (Kendall & Stewart, 1973),

coefficients based upon nonparametric alternatives to the t-test may provide indices that are unbiased in small samples and that are more statistically efficient than the usual indices. Statistics such as the rank biserial correlation (Glass, 1966: Cureton, 1968), or those used for nonparametric effect size estimations (Hedges & Olkin, 1984) should be explored for such applications.

The stability of the results obtained in this research across the ten examination forms provides evidence of the generalizability of the results. Unfortunately, the use of archival data files from operational tests as the populations from which samples were drawn imposed lower limits on the technical quality of the test items examined. For example, negatively discriminating items were almost entirely absent from the data files, such items having been eliminated or corrected during the test development process. Similarly, tests with marginal values of internal consistency (alphas of 0.5 or 0.6) were not available. Further research on traditional item analysis statistics, using test forms providing a greater range of values for these statistics, is needed to extend these results across the breadth of population values of the indices.

References

Conrad, S. H. (1948). Characteristics and uses of item-analysis data. Psychological Monographs, 62, 1-48.

Crocker, l. & Algina, J. (1986). Introductions to classical and modern test theory. New York: Holt, Rinehart, and Winston, Inc.

Cureton, E. E. (1968). Rank-biserial correlation when ties are present. Educational and Psychological Measurement, 28, 77-79.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Lien (Ed.), Educational measurement (pp. 147-220). Washington, DC: American Council Education.

Glass, G. V. (1966). Note on rank biserial correlation. Educational and Psychological Measurement, 26, 623-631.

Hedges, L. V., & Olkin, I. (1984). Nonparametric estimators of effect size in meta-analysis. Psychological Bulletin, 96-3, 73-380.

Henryson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, Dc: Council on Education.

Kendall, M. G., & Stewart, A. (1973). Advanced theory of Statistics. New York, N.Y.: Hasner Publishing Co.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Nevo, B. (1980). Item analysis with small samples. Applied Psychological Measurement, 4, 323-329.

Nunnally, J. C. (1967). Psychometric theory. New York: McGraw-Hill.

Perry, N. C., & Michael, W. B. (1954). A tabulation of the fiducial limits for the point-biserial correlation coefficient. Educational and Psychological Measurement, 14, 715-721.

Swineford, F. (1974). The test consultant manual. Princeton, NJ: Educational Testing Service.

Table 1

FTCE Subject Area Tests Used as Pseudo-populations

FTCE Subject Area Test	N of Items	N of Examinees	Total Score MN	SD
Biology	115	325	80.69	13.05
Elementary (1-6)	140	6405	93.47	12.88
Emotionally Handicapped	117	434	95.97	8.27
English (6-9)	83	594	62.57	8.24
Guidance	119	490	88.31	9.69
Mathematics (6-9)	96	635	55.49	14.33
Physical Education	119	390	76.64	11.90
Early Childhood (K-3)	141	1326	98.02	12.09
Social Studies	157	578	116.37	17.85
Specific Learning Disabilities	118	640	86.21	9.73

Figure 1

Distribution of the Statistical Biases in the Estimation of the Item Difficulty Index for Six Sample Sizes

Examination Form: 1

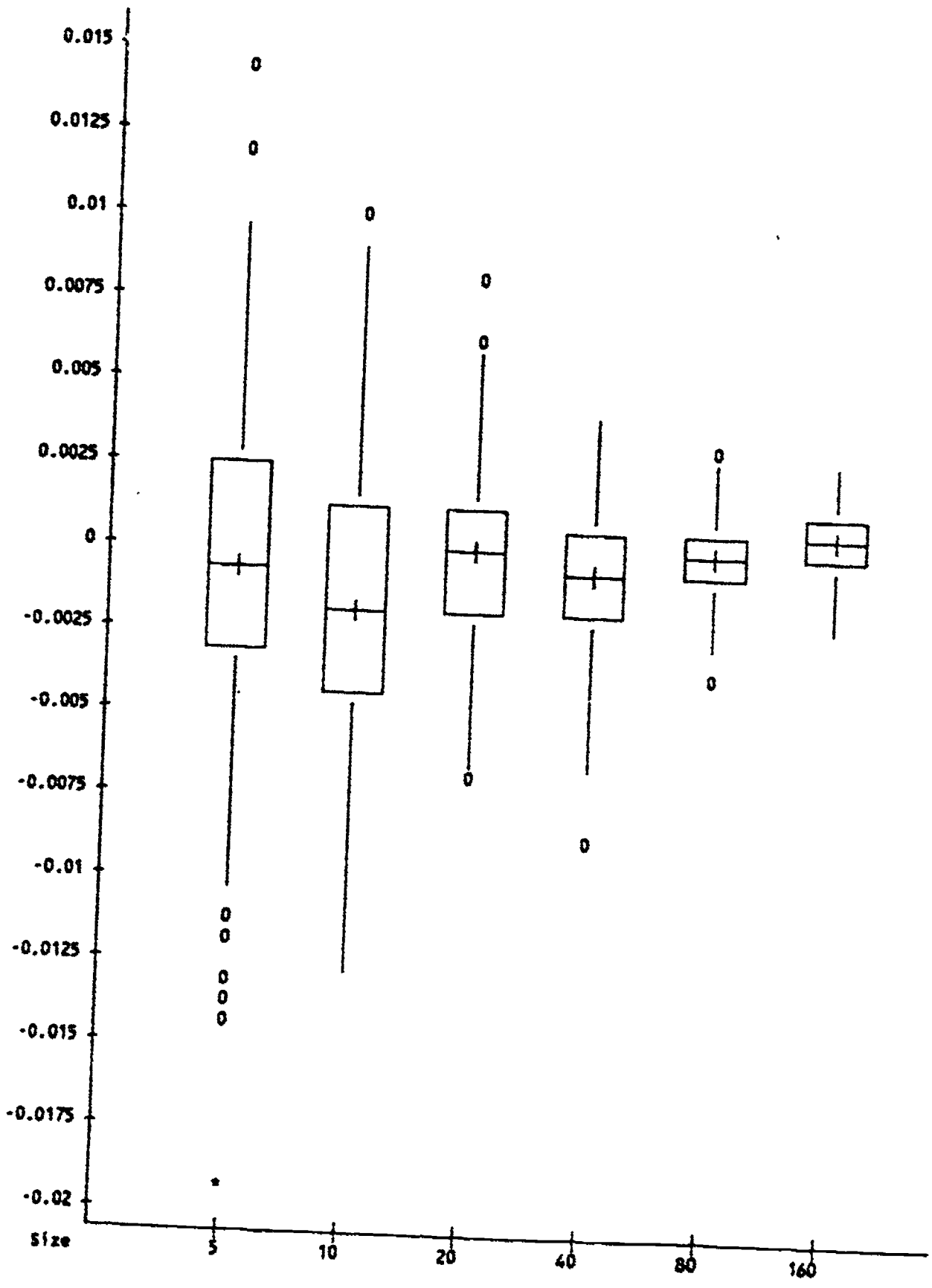

Table 2

Sampling Bias in the Estimation of the Item Difficulty Index

for Six Sample Sizes

Test Form: 1

	SAMPLE BIAS					
	SIZE					
	5	10	20	40	80	160
ITEM DIFFICULTY						
<.10	0.004	-0.001	0.001	-0.000	-0.001	0.001
.10-.19	-0.009	-0.000	0.002	0.002	-0.000	-0.001
.20-.29	0.001	0.002	-0.001	-0.002	-0.000	-0.000
.30-.39	0.000	-0.004	-0.000	-0.002	-0.000	0.001
.40-.49	0.000	-0.000	0.000	-0.001	-0.000	0.001
.50-.59	-0.004	-0.001	-0.001	-0.002	-0.000	0.001
.60-.69	0.000	-0.005	-0.001	-0.001	-0.000	0.001
.70-.79	-0.000	-0.003	0.000	-0.002	0.001	0.001
.80-.89	0.001	-0.002	0.000	-0.000	-0.000	0.000
.90-1.00	-0.001	-0.000	-0.000	-0.000	0.000	0.000

Table 3

Standard Error of the Estimate of the Item Difficulty Index

for Six Sample Sizes

Test Form: 1

	STANDARD ERROR					
	SIZE					
	5	10	20	40	80	160
ITEM DIFFICULTY						
<.10	0.134	0.089	0.065	0.045	0.033	0.023
.10-.19	0.155	0.112	0.080	0.055	0.040	0.028
.20-.29	0.194	0.143	0.101	0.069	0.049	0.035
.30-.39	0.212	0.153	0.107	0.075	0.054	0.038
.40-.49	0.220	0.157	0.111	0.077	0.055	0.039
.50-.59	0.222	0.157	0.112	0.078	0.055	0.040
.60-.69	0.214	0.153	0.107	0.076	0.054	0.038
.70-.79	0.196	0.140	0.098	0.070	0.049	0.035
.80-.89	0.167	0.120	0.084	0.059	0.042	0.030
.90-1.00	0.092	0.065	0.046	0.032	0.023	0.016

Figure 2
Standard Errors of Item Difficulty Index

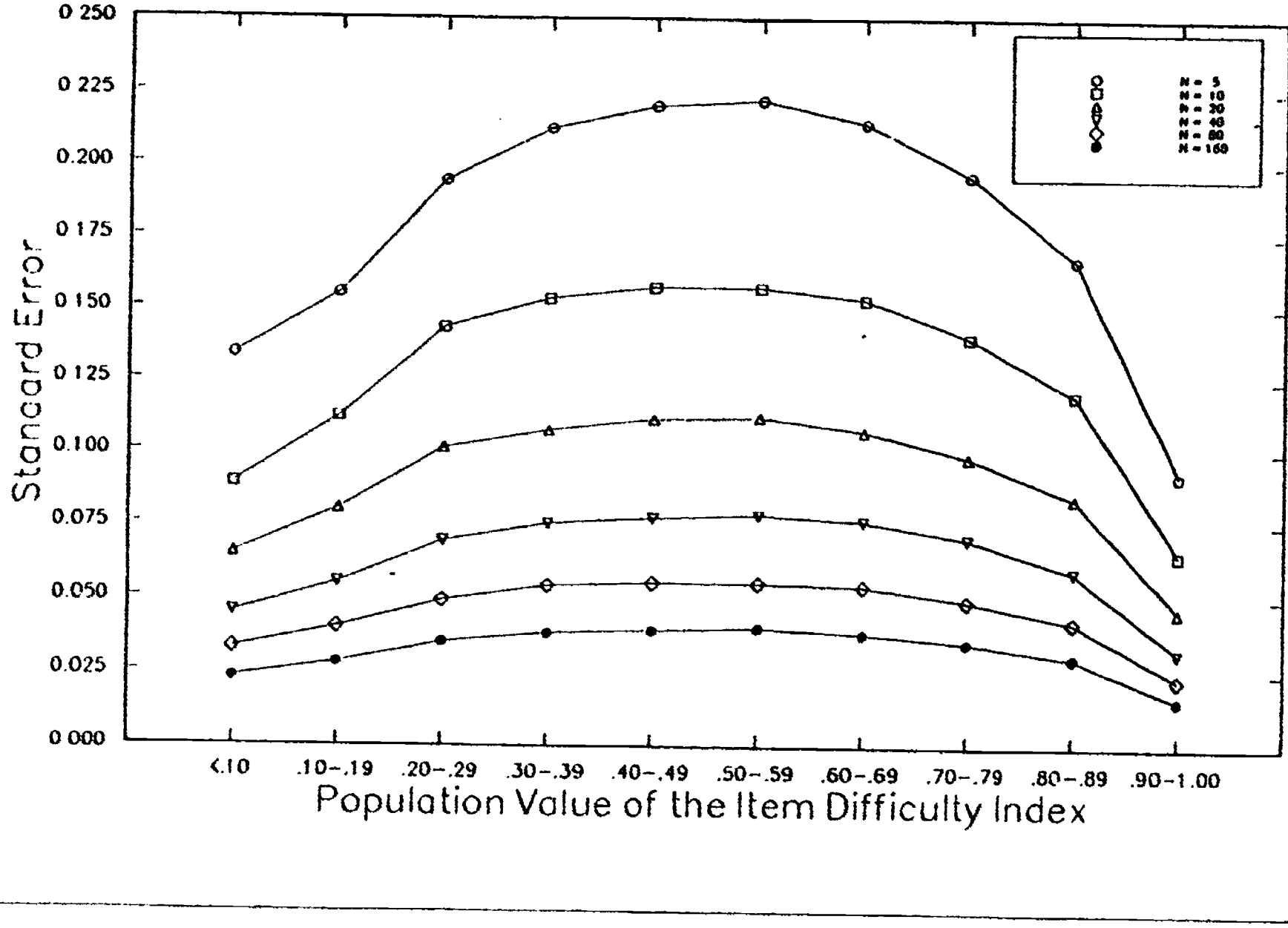

Figure 3

Distribution of the Statistical Biases in the Estimation of the Item Discrimination Index for Six Sample Sizes

Examination Form: 1

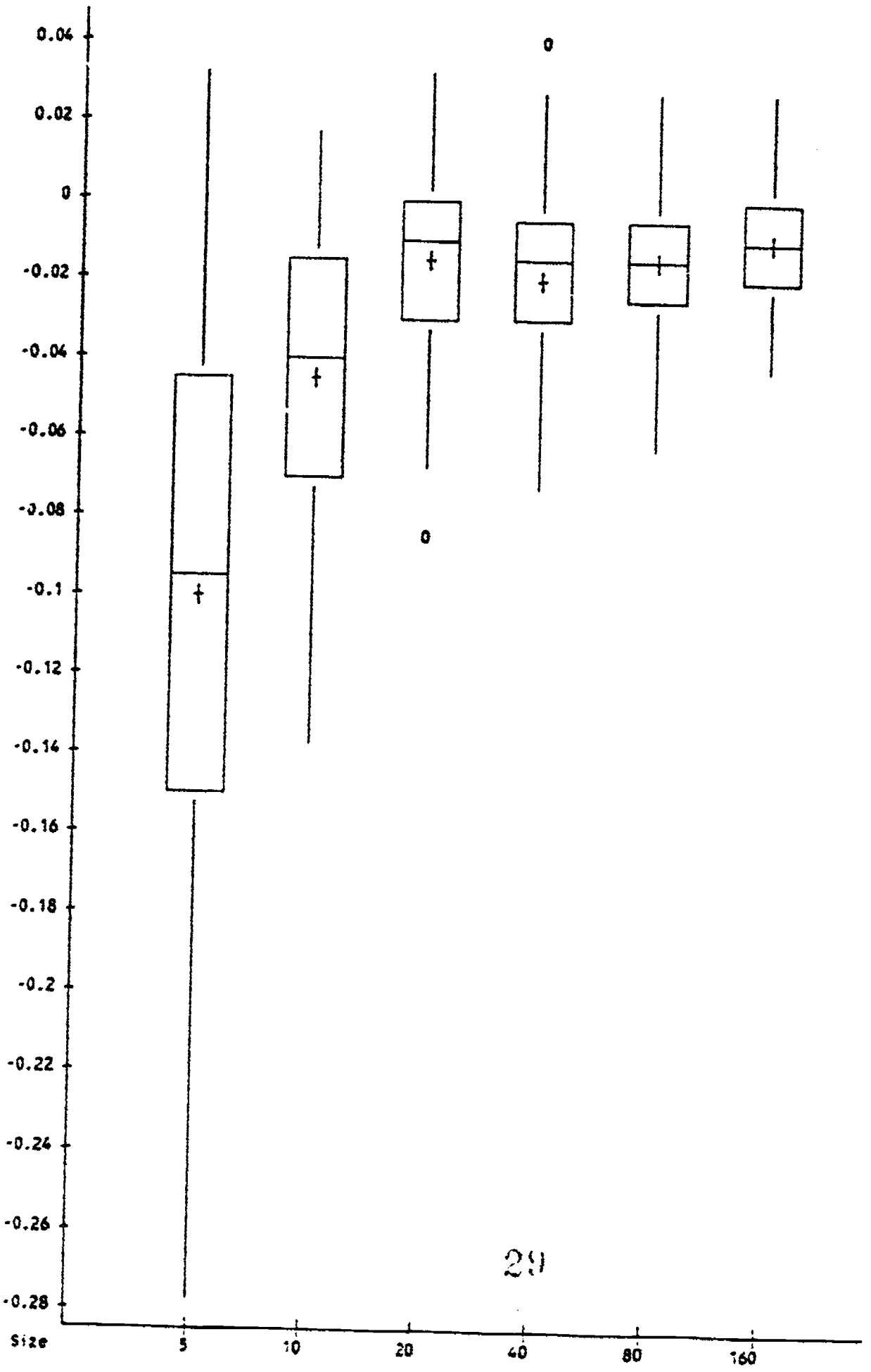

Table 4

Sampling Bias in the Estimation of the Item Discrimination Index

for Six Sample Sizes

Test Form: 1

	SAMPLE BIAS					
	SIZE					
	5	10	20	40	80	160
ITEM DISCRIMINATION						
<.10	-0.016	-0.004	-0.001	-0.003	-0.001	-0.001
.10-.19	-0.051	-0.023	-0.008	-0.011	-0.011	-0.007
.20-.29	-0.091	-0.035	-0.011	-0.016	-0.014	-0.010
.30-.39	-0.117	-0.053	-0.012	-0.020	-0.016	-0.014
.40-.49	-0.152	-0.075	-0.032	-0.033	-0.027	-0.021
.50-.59	-0.180	-0.078	-0.022	-0.026	-0.018	-0.012
.60-.69	-0.219	-0.119	-0.039	-0.035	-0.027	-0.018
.70-.79	-0.242	-0.100	-0.039	-0.041	-0.028	-0.018

Figure 4
Bias in the Estimation of the Item Discrimination Index

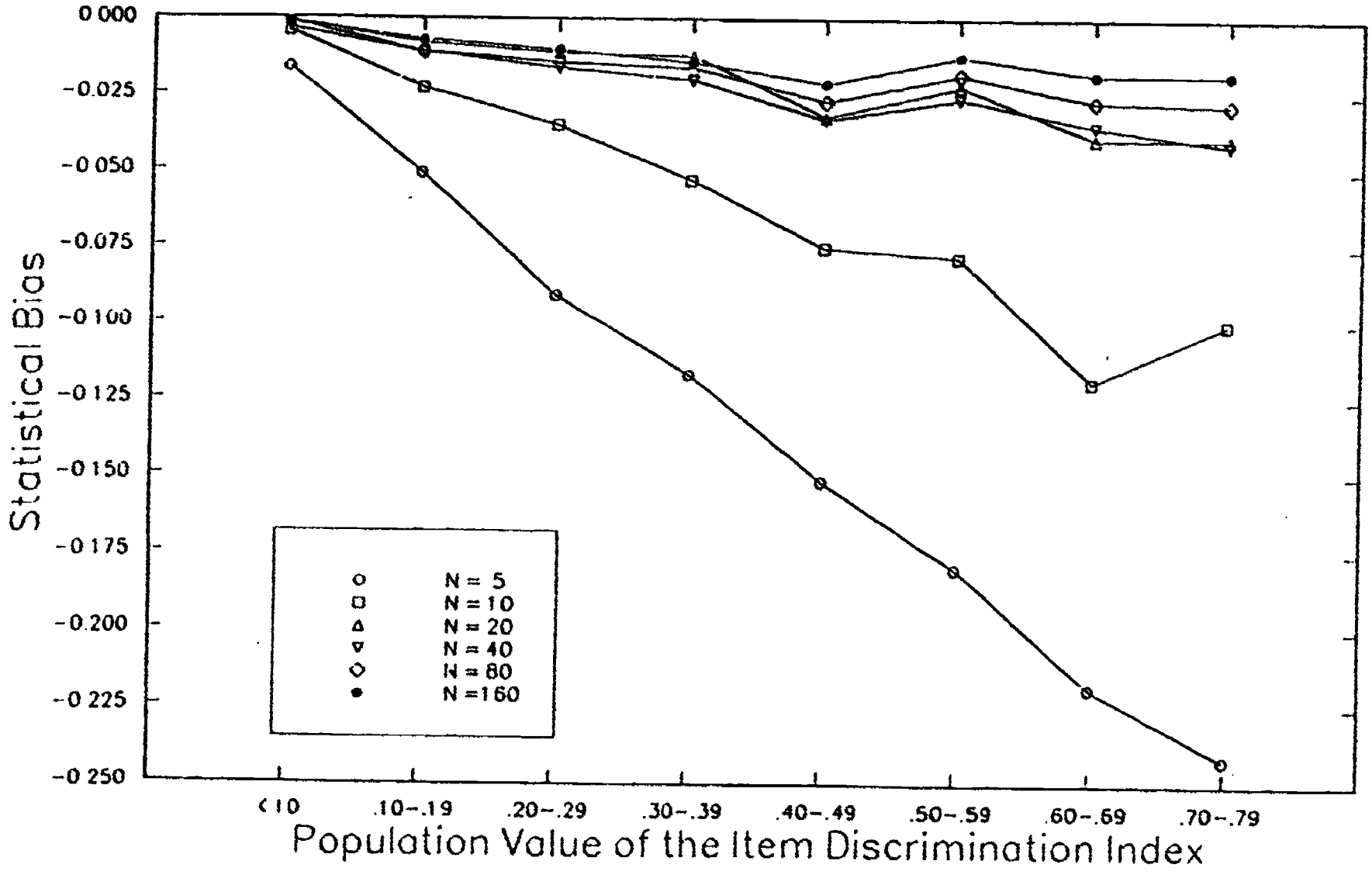

Table 5

Standard Error of the Estimate of the Item Discrimination Index

for Six Sample Sizes

Test Form: 1

	STANDARD ERROR					
	SIZE					
	5	10	20	40	80	160
ITEM DISCRIMINATION						
<.10	0.273	0.197	0.155	0.108	0.077	0.052
.10-.19	0.386	0.292	0.215	0.143	0.098	0.074
.20-.29	0.434	0.349	0.268	0.182	0.130	0.093
.30-.39	0.447	0.351	0.272	0.184	0.130	0.094
.40-.49	0.426	0.359	0.277	0.185	0.132	0.094
.50-.59	0.412	0.333	0.266	0.175	0.128	0.090
.60-.69	.	0.301	0.237	0.162	0.119	0.087
.70-.79	.	.	0.222	0.151	0.108	0.077

Figure 5
Standard Errors of Item Discrimination Index

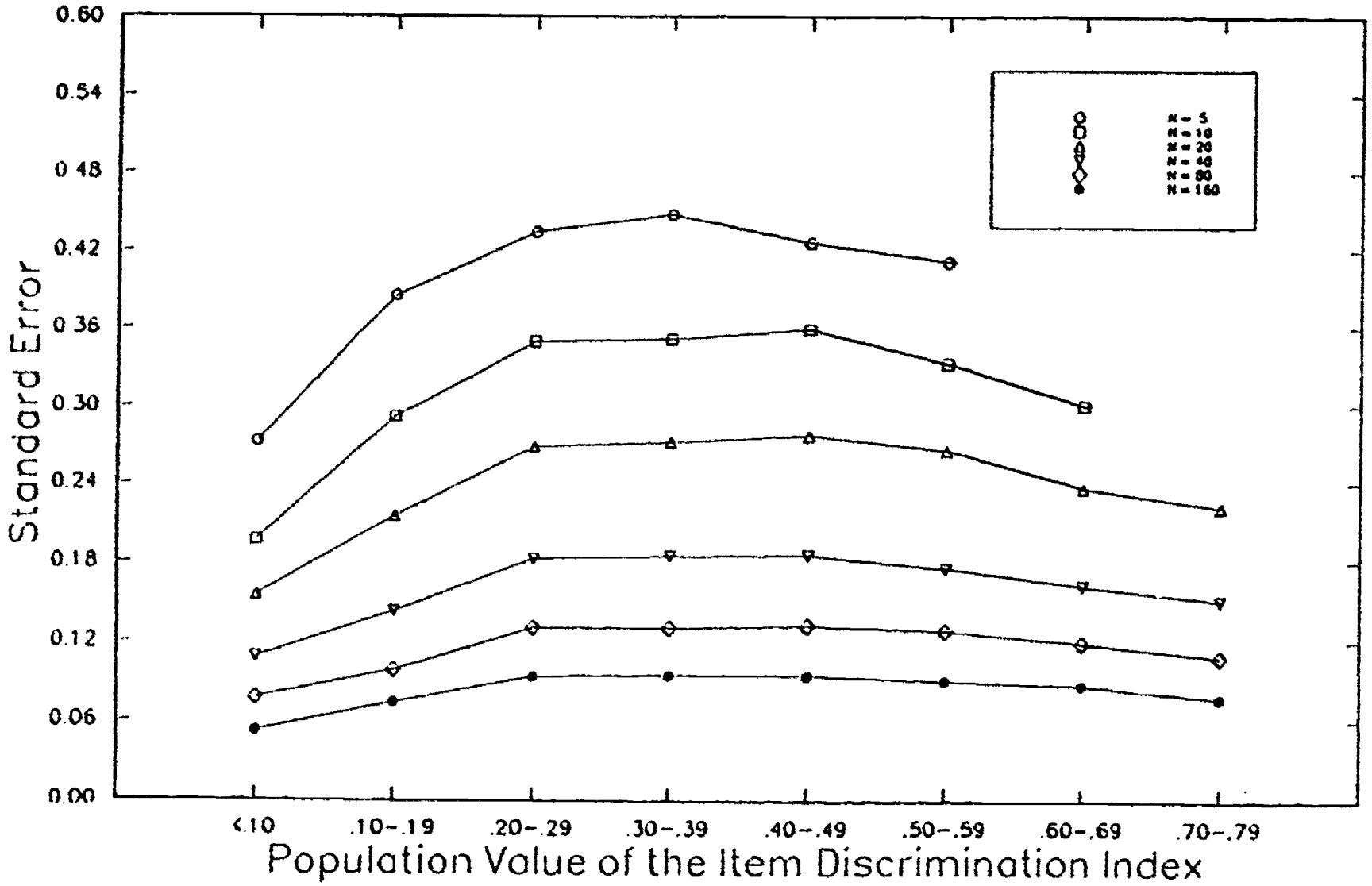

Figure 6

Distribution of the Statistical Biases in the Estimation

of the Item-Total Point-Biserial Correlation for Six Sample Sizes

Examination Form: 1

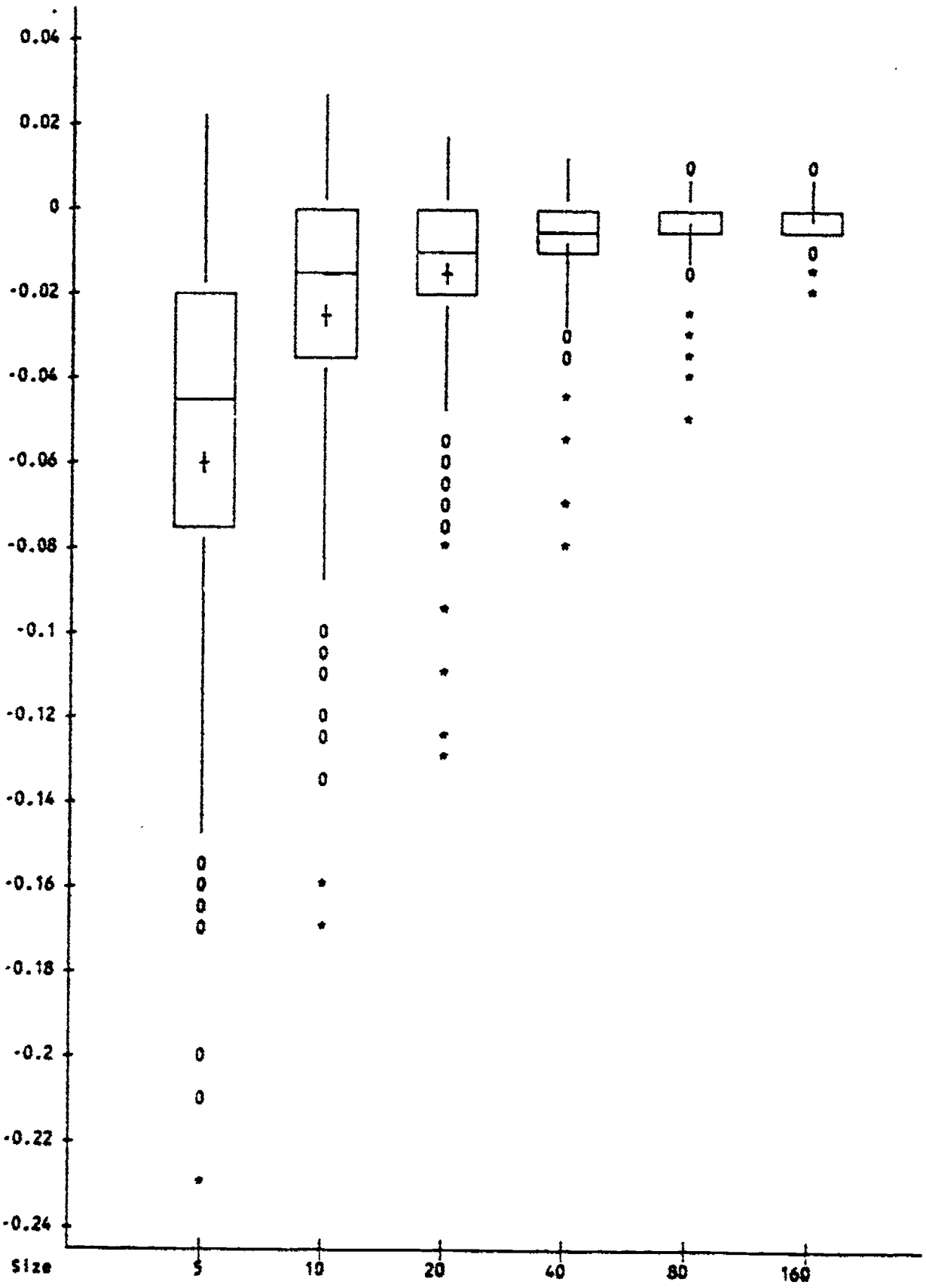

Table 6

Sampling Bias in the Estimation

of the Item-Total Point-Biserial Correlation Index

for Six Sample Sizes

Test Form: 1

	SAMPLE BIAS					
	SIZE					
	5	10	20	40	80	160
POINT BISERIAL						
<.10	-0.020	-0.010	-0.008	-0.005	-0.004	-0.000
.10-.19	-0.051	-0.025	-0.019	-0.011	-0.006	-0.002
.20-.29	-0.058	-0.030	-0.017	-0.009	-0.004	-0.002
.30-.39	-0.080	-0.037	-0.021	-0.010	-0.004	-0.002
.40-.49	-0.064	-0.028	-0.014	-0.002	-0.000	-0.002
.50-.59	-0.056	-0.009	-0.003	0.001	0.001	0.000

Figure 7

Bias in the Estimation of the Point–Biserial Correlation

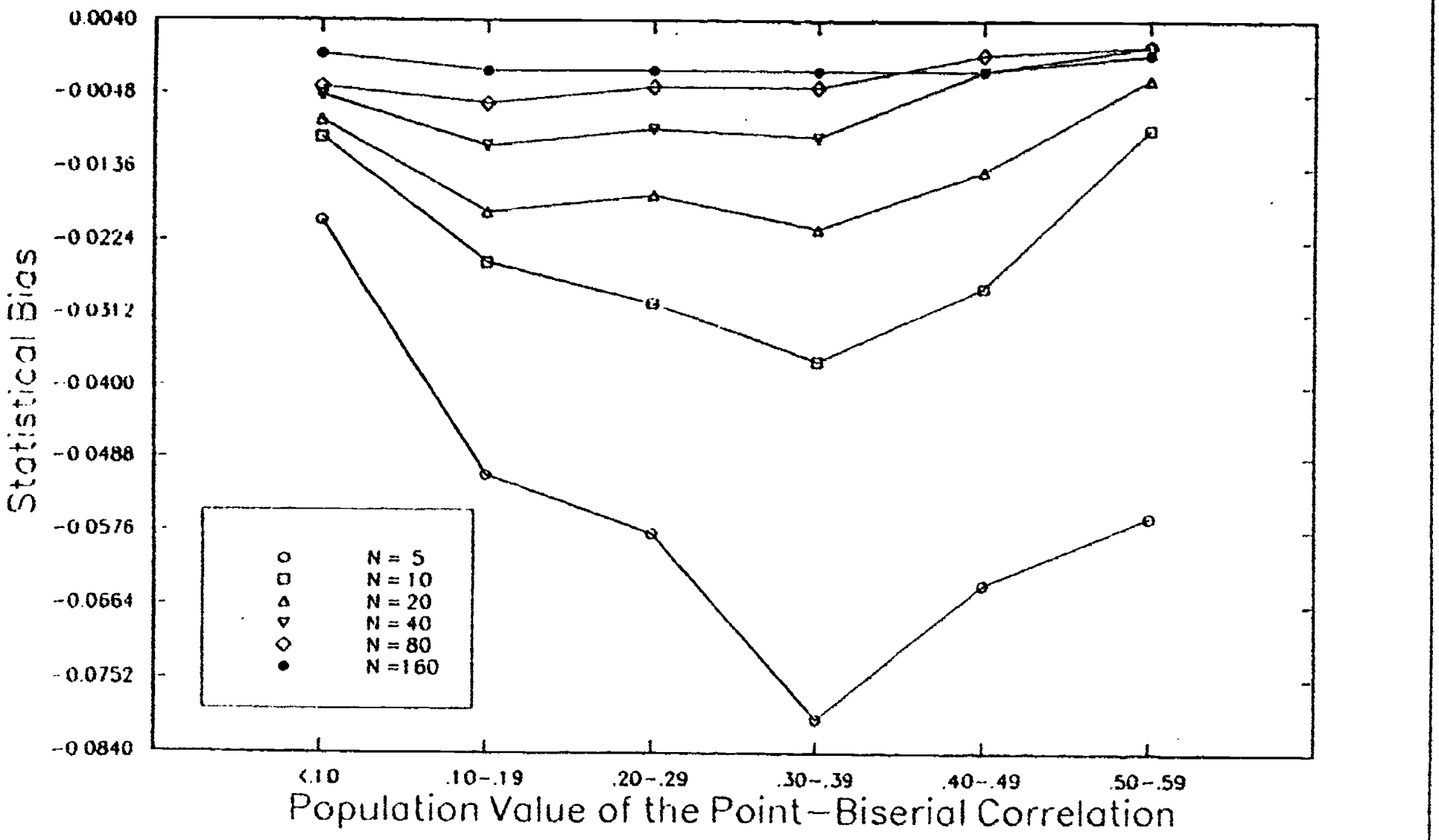

Table 7

Standard Error of the Estimate

of the Item-Total Point-Biserial Correlation

for Six Sample Sizes

Test Form: 1

	STANDARD ERROR					
	SIZE					
	5	10	20	40	80	160
POINT BISERIAL						
<.10	0.305	0.235	0.186	0.141	0.100	0.070
.10-.19	0.407	0.301	0.229	0.168	0.118	0.083
.20-.29	0.428	0.310	0.222	0.158	0.112	0.080
.30-.39	0.427	0.303	0.215	0.146	0.103	0.073
.40-.49	0.405	0.278	0.195	0.134	0.095	0.067
.50-.59	0.379	0.239	0.165	0.112	0.080	0.056

Figure 8
Standard Errors of Point–Biserial Correlation

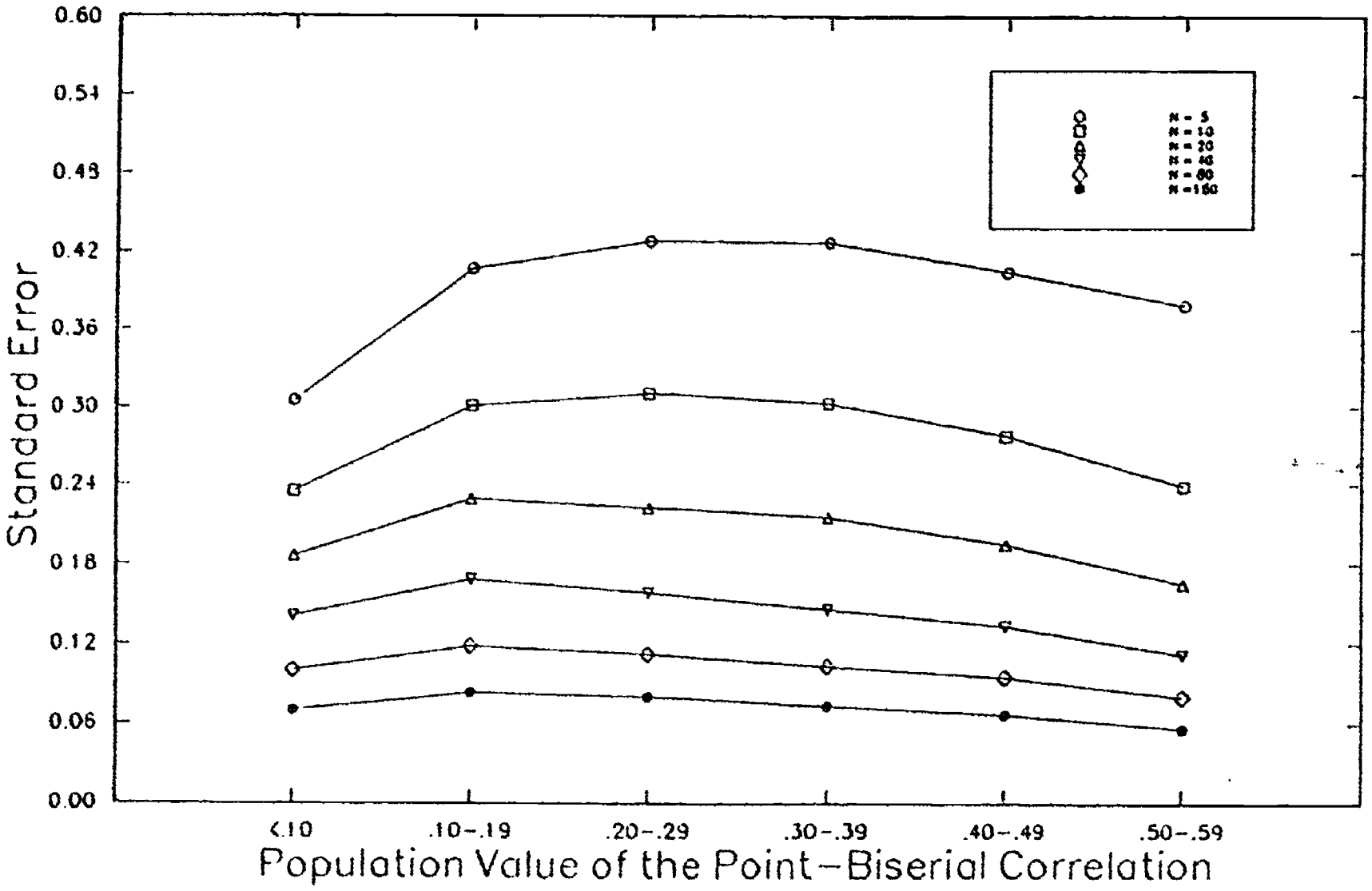

Table 8

Sample Bias in Estimation of Coefficient Alpha By Test Form and Sample Size

Sample Size	Test Form									
	1	2	3	4	5	6	7	8	9	10
5	-0.070	-0.096	-0.126	-0.101	-0.104	-0.063	-0.090	-0.102	-0.059	-0.090
10	-0.022	-0.053	-0.091	-0.061	-0.067	-0.019	-0.044	-0.063	-0.019	-0.061
20	-0.012	-0.028	-0.046	-0.035	-0.034	-0.009	-0.020	-0.032	-0.010	-0.025
40	-0.005	-0.010	-0.023	-0.016	-0.019	-0.004	-0.006	-0.012	-0.004	-0.014
80	-0.002	-0.008	-0.013	-0.008	-0.008	-0.002	-0.005	-0.007	-0.001	-0.006
160	-0.001	-0.002	-0.005	-0.005	-0.004	0.000	-0.002	-0.002	-0.001	-0.003

Table 9

Standard Errors of Coefficient Alpha for Six Sample Sizes

Sample Size	Test Form									
	1	2	3	4	5	6	7	8	9	10
5	0.157	0.184	0.217	0.207	0.203	0.142	0.187	0.197	0.138	0.196
10	0.064	0.129	0.169	0.146	0.150	0.058	0.124	0.154	0.053	0.139
20	0.043	0.078	0.107	0.097	0.099	0.035	0.075	0.091	0.030	0.081
40	0.025	0.042	0.074	0.057	0.062	0.023	0.042	0.053	0.018	0.054
80	0.018	0.032	0.048	0.038	0.040	0.015	0.029	0.036	0.012	0.035
160	0.011	0.021	0.034	0.026	0.029	0.011	0.020	0.025	0.008	0.025

Figure 9

Distribution of Sample Estimates of Coefficient Alpha for Six Sample Sizes

Examination Form: 1